DOCUMENT RESUME

ED 424 237 TM 027 363

AUTHOR Glas, Cees A. W.
TITLE Testing the Generalized Partial Credit Model. Research Report 96-03.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE 1996-10-00
NOTE 45p.
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Foreign Countries; *Item Response Theory; *Mathematical Models; *Statistical Analysis
IDENTIFIERS Dimensionality (Tests); *Partial Credit Model; Power (Statistics); *Rasch Model

ABSTRACT

The partial credit model (PCM) (G. N. Masters, 1982) can be viewed as a generalization of the Rasch model for dichotomous items to the case of polytomous items. In many cases, the PCM is too restrictive to fit the data. Several generalizations of the PCM have been proposed. In this paper, a generalization of the PCM (GPCM), a further generalization of the one-parameter logistic model, is discussed. The model is defined and the conditional maximum likelihood procedure for the method is described. Two statistical tests for the model, based on generalized Pearson statistics, are presented. The first is a generalization of some well-known statistics for the Rasch model for dichotomous items to the GPCM which has power against incorrect specifications of the form of the item characteristic curves. The other test has power against local dependence and multidimensionality, and is built on an approach introduced by A. L. van den Wollenberg (1982) and C. A. W. Glas (1988) for testing unidimensionality in the Rasch model for dichotomous items. Some simulation studies are presented concerning the power of the tests. (Contains 31 references.) (SLD)

TM

# Testing the Generalized Partial Credit Model

**Research Report 96-03**





Cees A.W. Glas

# Testing the Generalized Partial Credit Model

Cees A.W. Glas

Testing the Generalized Partial Credit Model, Cees A.W. Glas - Enschede: University of Twente, Faculty of Educational Science and Technology, December 1996. - 41 pages.

## Introduction

The partial credit model (Masters, 1982) can be viewed as a generalization of the Rasch model for dichotomous items (Rasch, 1960) to the case of polytomous items. As the Rasch model, the partial credit model has desirable mathematical properties, which arise from the fact that the model defines an exponential family. The major advantage of an exponential family IRT model is that there exist minimal sufficient statistics for both the item and person parameters. Conditioning on the sufficient statistics for the ability parameters facilitates so-called conditional maximum likelihood (CML) estimation of item parameters, which has the advantage that no assumptions about the distribution of ability have to be made and that, from the point of view of parameter estimation, random sampling of respondents is not necessary (Rasch, 1960; Andersen, 1972). Also the conditional likelihood defines an exponential family, which allows for a relatively simple estimation procedure, where the minimal sufficient statistics are equated with their expected values. Further, except for certain boundary values of the sufficient statistics, there exists a unique solution to the estimation equations. Another consequence of the favorable mathematical structure of the model is the possibility to develop proper statistical testing procedures (Andersen, 1973; Martin Löf, 1973; Glas, 1988), that is, testing procedures based on statistics with a proven (asymptotic) distribution, which are informative with respect to specific model violations. However, in many instances, the partial credit model (PCM) is too restrictive to fit the data. Therefore, several generalizations of the PCM have been proposed. Bock (1972), for instance, introduces discrimination parameters for the item categories. Therefore, this generalization is comparable to the generalization of the Rasch model for dichotomous items to the two-parameter model (Birnbaum,

1968). However, as with the two-parameter model, also Bock's PCM no longer defines an exponential family, and CML estimation is no longer possible. Marginal maximum likelihood (MML) estimation (Bock and Aitkin, 1981), where the model is extended with assumptions concerning an ability distribution, seems to solve some problems, but with respect to testing procedures based on statistics with known (asymptotic) distributions, little progress has been made.

Another generalization of the PCM is the One Parameter Logistic Model (OPLM, Verhelst & Glas, 1995; Verhelst, Glas & Verstralen, 1995). Here, for every item, a discrimination index is imputed as a known constant and only the item difficulty parameters are estimated. By imputing and not estimating discrimination indices, OPLM, unlike the two-parameter logistic model, preserves the powerful mathematical properties of exponential family models. Further, Glas and Verhelst (1995) have developed a method where hypothesis concerning the magnitude of the discrimination indices are iteratively defined and tested until a possible model fit is obtained.

The version of the PCM by Wilson and Masters (1993) can be seen as a further generalization of the OPLM, although was it developed from an entirely different point of view. This model was first motivated by the problem that item parameters in the PCM cannot be estimated if certain response categories are unobserved. The idea is as follows. Suppose that an item has 5 response categories {0,1,2,3,4}, and the third category is not responded to. Then the item format is transformed to 4 categories with weights {0,1,3,4}. If the first category is unobserved, the category weights will be {1,2,3,4}. In this way, the relative contribution of the various items to the sufficient statistic for ability, that is, the sum score, is not altered by the presence of unobserved categories. However, the model can also be seen as a further generalization of the OPLM, where the

scoring weights associated with the categories account for the differences in discrimination between the categories within items. The rest of this chapter will be devoted to this generalization of the PCM, abbreviated GPCM. First, some preliminaries will be given: a formal definition of the model and the CML procedure. Next, two statistical tests for the model will be discussed. These tests are based on generalized Pearson statistics (Glas and Verhelst, 1989, 1995). The first is a generalization of some well-known statistics for the Rasch model for dichotomous items (Martin Löf, 1973; Glas, 1988) to the GPCM and has power against incorrect specification of the form of the item characteristic curves. The theory of these tests is worked out in Glas and Verhelst (1995), where CML and MML estimation and testing are described for a general model that includes the GPCM as a special case. This test is discussed in this chapter to contrast it with a new test to be presented here, that has power against local dependence and multidimensionality. This last test is built on an approach introduced by van den Wollenberg (1982) and Glas (1988) for testing unidimensionality in the Rasch model for dichotomous items. This chapter will be concluded with some simulation studies concerning the power of the tests.

## The Model

Let item $i$ have $m_i+1$ response categories indexed $h = 0,1,...,m_i$. The response to the item will be represented by an $(m_i+1)$-dimensional vector $\mathbf{x}_i'=(x_{i0},...,x_{ih},...,x_{im_i})$, where $x_{ih}$ is defined by

$$x_{ih} = \begin{cases} 1 & \text{if a response is given in category } h, \\ 0 & \text{if this is not the case.} \end{cases} \quad (1)$$

The probability of a response in category $h$, $h = 0,...,m_i$ as a function of an ability parameter $\theta$ and a vector of the parameters of item $i$, $\beta_i' = (\beta_{i1},...,\beta_{ih},...,\beta_{im_i})$, is given by

$$P(X_{ih} = 1 | \theta, \beta_i) = \frac{\exp(r_{ih}\theta - \sum_{g=1}^{h} \beta_{ig})}{\sum_{g=0}^{m_i} \exp(r_{ig}\theta - \sum_{k=1}^{g} \beta_{ik})}, \quad (2)$$

where the summation $\sum_{k=1}^{0}$ is supposed to have a zero result. As with the usual PCM, the parameters $\beta_{ih}$ are the values on the $\theta$-scale where $P(X_{ih} = 1 | \theta, \beta_i)$ and $P(X_{i,h-1} = 1 | \theta, \beta_i)$ are equal. Although this parameterization of the model entails a nice interpretation of the parameters, for mathematical reasons it is convenient to introduce the reparametrization $\eta_{ih} = \sum_{g=1}^{h} \beta_{ig}$, $h = 1,...,m_i$, and write the model as

$$P(X_{ih}=1 | \theta, \eta_i) = \frac{\exp(r_{ih}\theta - \eta_{ih})}{\sum_{g=0}^{m_i} \exp(r_{ig}\theta - \eta_{ig})}, \quad (3)$$

where $\eta_i$ has elements $\eta_{ih}$, for $h = 0,...,m_i$ and $\eta_{i0}$ is fixed to zero. Introducing $\boldsymbol{r}_i' = (r_{i0},...,r_{ih},...,r_{im_i})$ the probability of response $\boldsymbol{x}_i$ can be written as

$$P(\boldsymbol{x}_i | \theta, \eta_i) \propto \exp(\boldsymbol{x}_i'(\boldsymbol{r}_i\theta - \eta_i)). \quad (4)$$

As a concluding remark in this section, the following is in order. Notice that in the parametrization of (2) and (3), it is possible to have an item with, say, $m_i = 2$ and

score weights {1, 2, 3}, that is, the zero score cannot be obtained on this item. For practical purposes, such as not having to down-code data in case of a missing zero category, and for communication of results to the practitioner, this may be quite convenient and all theory to be presented below applies to the general parametrization (1) and (2). However, it must be stressed that subtracting a weight equal to $r_{i0}$ from all category weights within the item, such that $r_{i0}$ itself will be transformed to zero, will not alter the likelihood equations. With this alteration the denominator of (3) will equal $1 + \sum_{g=1}^{m_i} \exp(r_{ig}\theta - \eta_{ig})$, , while the nominator of the probability of scoring in the zero category will equal one.

## CML Estimation

For deriving the asymptotic distribution of the statistics to be presented below, consistency of parameter estimates is essential. However, in the Rasch model, the number of person parameters goes to infinity if the sample size goes to infinity, and it is well known that, in general, this results in inconsistency of the maximum likelihood estimates (Neyman & Scott, 1948). In the psychometric literature, two ways are suggested to get rid of the person parameters, maximizing a conditional likelihood which only depends on the item parameters (CML, Rasch, 1960, 1961, Andersen, 1973, 1977) and maximizing the likelihood function of a model extended with an ability distribution (MML, Bock & Aitkin, 1981). Both approaches can be applied to estimating the GPCM. However, since it is based on fewer assumptions, in general, CML is the preferable estimation method, and therefore only CML will be worked out here.

In the derivations of the asymptotic distribution of the below statistics, the

general framework of parameterized multinomial models is used. This framework will also be used for describing the essentials of CML estimation. Let a test consist of $K$ items and let $\boldsymbol{x}$ be a response pattern, so $\boldsymbol{x}' = (\boldsymbol{x}_1', ..., \boldsymbol{x}_i', ..., \boldsymbol{x}_K')$.. In this framework the data are viewed as counts $n_{\boldsymbol{x}}$, for all $\boldsymbol{x} \in \{\boldsymbol{x}\}$, which is the set of all possible response patterns. The number of possible response patterns will be denoted by $M$. In a CML framework, the probabilities of the response patterns are derived as follows. Using (4) and local stochastic independence, the probability of response pattern $\boldsymbol{x}$ as a function of the ability and item parameters is given by

$$P(\boldsymbol{x}|\theta,\eta) \propto \exp(\boldsymbol{x}'(\boldsymbol{r}\theta - \eta)), \tag{5}$$

where $\boldsymbol{r}' = (\boldsymbol{r}_1', ..., \boldsymbol{r}_i', ..., \boldsymbol{r}_K')$ and $\eta' = (\eta_1', ..., \eta_i', ..., \eta_K')$.

For all possible outcomes $\boldsymbol{x}$, a sufficient statistic $s$ is defined by $s = \boldsymbol{x}'\boldsymbol{r}$ and for every possible $s$, a set $\{\boldsymbol{x}|s = \boldsymbol{x}'\boldsymbol{r}\}$ is defined. Notice that $\bigcup_S \{\boldsymbol{x}|s = \boldsymbol{x}'\boldsymbol{r}\} = \{\boldsymbol{x}\}$. The conditional probability of response pattern $\boldsymbol{x}$ given the associated value of $s$ is given by

$$\begin{aligned} P(\boldsymbol{x}|s,\eta) &= \frac{\exp(\boldsymbol{x}'(\boldsymbol{r}\theta - \eta))}{\sum_{\{\boldsymbol{y}|\boldsymbol{y}'\boldsymbol{r}=s\}} \exp(\boldsymbol{y}'(\boldsymbol{r}\theta - \eta))} \\ &= \frac{\exp(-\boldsymbol{x}'\eta)}{\sum_{\{\boldsymbol{y}|\boldsymbol{y}'\boldsymbol{r}=s\}} \exp(-\boldsymbol{y}'\eta)} \\ &= \frac{\exp(-\boldsymbol{x}'\eta)}{\gamma_s}, \end{aligned} \tag{6}$$

where $\gamma_s$ is a combinatorial function defined by

$$\gamma_S = \sum_{\{y \mid y'r = s\}} \exp(-y'\eta). \tag{7}$$

Notice that these probabilities are a function of the item parameters only. Maximizing the likelihood function associated with this model produces the desired CML estimates.

The model defined by (6) does not yet fit the framework of the multinomial model: the probabilities of the response patterns resulting in the same value s sum to one, and as a consequence, the distribution function of the counts of the response patterns has a product-multinomial form. This problem is easily solved using a well-known procedure by Birch (1963, see also Haberman, 1974). Let $\{s\}$ be the set of all possible values of $s$. For all $s \in \{s\}$, let $N_S$ be the number of persons in the sample obtaining a score $s$. Assume that $N_S, s \in \{s\}$, has a multinomial distribution, indexed by the sample size $N$ and the parameters $\omega_S$ for all $s \in \{s\}$. Notice that the ML estimate of $\omega_S$ is given by $\hat{\omega}_S = n_S/N$. Using (6), the probability of response pattern $\boldsymbol{x}$ can now be given by

$$\pi_{\boldsymbol{x}} = P(\boldsymbol{x} \mid \omega, \eta) = \frac{\omega_S \exp(-\boldsymbol{x}'\eta)}{\gamma_S}, \tag{8}$$

with $\omega$ a vector with elements $\omega_S$ for all $s \in \{s\}$. It is easily verified that the probabilities (8) sum to one, so the model now fits the general multinomial model.

Next, it will be shown that (8) defines an exponential family. A model belongs to the exponential family if the likelihood function of parameters $\phi$ given an observation $\boldsymbol{x}$ can be written as

$$L(\phi; \boldsymbol{x}) = c(\boldsymbol{x}) \exp(\phi' t(\boldsymbol{x}))/a(\phi), \tag{9}$$

where $\mathbf{t}(\boldsymbol{x})$ is a vector of functions of $\boldsymbol{x}$, and $c(.)$ and $a(.)$ are functions only of $\boldsymbol{x}$ and $\phi$, respectively. The likelihood of the Rasch model, enhanced with a saturated multinomial model for the score distribution can be written as

$$L(\eta,\omega;\boldsymbol{x}) = \frac{\exp(\sum_{i,j} x_{ij}\eta_{ij}+\sum_j \delta_{js}\ln\omega_j)}{\gamma_s(\varepsilon)}, \qquad (10)$$

where $\delta_{js}$ is the Kronecker delta, taking the value one if $j = s$ and zero otherwise. Comparing (9) and (10) it can easily be verified that (10) does indeed define an exponential family. Notice that the restriction $\sum_s \omega_s = 1$ implies that there are $|\{s\}|-1$ free parameters for the saturated model for the frequency distribution of the respondent's sum scores. Since also the item parameters need a restriction to produce a unique solution of the likelihood equations, the total number of free parameters $F$ is equal to $\sum_i m_i + |\{s\}| - 2$. Let $\boldsymbol{T}$ be defined as an $M\times(F+2)$ matrix which, for the $M$ different patterns, has the sufficient statistics $\mathbf{t}(\boldsymbol{x})'$, defined in (9), as rows. The rows are in an arbitrary but fixed order. The matrix $\boldsymbol{T}$ will be partitioned $(\boldsymbol{T}_1|\boldsymbol{T}_2)$. The matrix $\boldsymbol{T}_1$ has $\sum_i m_i$ columns, each column corresponding to an item parameter, the matrix $\boldsymbol{T}_2$ has $|\{s\}|$ columns, each column corresponding to a score. An example of the $\boldsymbol{T}$-matrix for 3 items is given in Table 1. The items have the score weights {1, 2}, {0, 1, 3} and {0, 1, 2, 3}, respectively. For convenience, the scores of the response patterns, the sufficient statistics, are given in the first column of Table 1. Using (10), the reader can verify that, for any response pattern $\boldsymbol{x}$, $\boldsymbol{T}_1$ has a row $\boldsymbol{x}$ and the columns of $\boldsymbol{T}_2$ are indicator vectors of the scores.

---

Insert Table 1 about here

---

It is well-known (see, for instance, Andersen, 1980) that in exponential family models, ML estimation boils down to equating the realizations of the minimal sufficient statistics to their expected values. So the likelihood equations can be written as

$$\boldsymbol{T}\boldsymbol{p} = \boldsymbol{T}\pi\,, \tag{11}$$

where $\boldsymbol{p}$ and $\pi$ are $M$-vectors with elements $p_{\boldsymbol{x}}$, the proportion respondents with response pattern $\boldsymbol{x}$ and the probability $\pi_{\boldsymbol{x}}$ defined by (8), respectively. Since $\boldsymbol{T}$ is related to an over-parameterization of the model, solving (11) will need two restrictions, one on the item parameters and one to assure that the dummy parameters $\omega_s$ sum to one. However, in the sequel it will become clear that considering a matrix $\boldsymbol{T}$ associated with an over-parametrization will prove convenient for the introduction of the test statistics.

## Generalized Pearson Tests

Let $\hat{\pi}$ be the vector $\pi$ evaluated using a BAN estimate (Best Asymptotically Normal estimate, say an ML estimate) of $\pi$. Further, $\boldsymbol{D}_{\pi}$ is the diagonal $M\times M$ matrix of the elements of $\pi$, and a vector of deviates $\boldsymbol{b}$ is defined as

$$\boldsymbol{b} = N^{1/2} (\boldsymbol{p} - \hat{\pi}). \tag{12}$$

The tests considered here, will be based on a vector of $G$ linear combinations $\boldsymbol{d} = \boldsymbol{U}'\boldsymbol{b}$, where the $M \times G$ matrix $\boldsymbol{U}$ is chosen such that $G < M$ and the linear combinations may show specific model violations. A method for constructing the so-called the matrix of contrasts $\boldsymbol{U}$ will be discussed below. Consider the statistic

$$Q = Q(\boldsymbol{U}) = \boldsymbol{b}'\boldsymbol{U}(\boldsymbol{U}'\boldsymbol{D}_{\pi}^{-1}\boldsymbol{U})^{-}\boldsymbol{U}'\boldsymbol{b} = \boldsymbol{d}'\boldsymbol{W}^{-}\boldsymbol{d}, \tag{13}$$

where $(\boldsymbol{U}\boldsymbol{D}_{\pi}\boldsymbol{U})^{-}$ and $\boldsymbol{W}^{-}$ stand for the generalized inverse of $(\boldsymbol{U}\boldsymbol{D}_{\pi}\boldsymbol{U})$ and $\boldsymbol{W}$, respectively. In the sequel, the vector $\boldsymbol{d}$ and the matrix $\boldsymbol{W}$ will be called the vector of deviates and the matrix of weights.

Glas and Verhelst (1989) have derived sufficient conditions for $Q$ to be asymptotically chi-square distributed with degrees of freedom equal to column-rank($\boldsymbol{U}$) - column-rank($\boldsymbol{T}$) - 1, or column-rank($\boldsymbol{U}$) - $F$ - 1. The general condition and the proof are beyond the scope of the present chapter. Here, only a method of constructing test statistics for exponential family models will be presented which guarantees that the conditions are satisfied and the asymptotic chi-square distribution is established.

Consider a matrix $\boldsymbol{U}$ that can be partitioned $\boldsymbol{U} = (\boldsymbol{T}|\boldsymbol{Y})$. Here $\boldsymbol{T}$ is a matrix as defined in the previous section and $\boldsymbol{Y}'\boldsymbol{b}$ is the matrix that produces the observed and expected frequencies of interest. In the example of the previous section, $\boldsymbol{T}$ was associated with a very specific over-parametrization of the GPCM. For exponential families in general, the over-parametrization has to be such that an $M$-vector with all elements equal to unity belongs to the manifold of $\boldsymbol{T}$. Obviously,

this is the case for the over-parametrization for the GPCM, since adding all columns of $T_2$ produces the desired $M$-vector.

For the reason for this restriction, one is referred to Glas (1988, 1989), Glas and Verhelst (1989) and Verhelst and Glas (1995). Some of its background will be commented upon in the sequel.

Let $\boldsymbol{d}' = (\boldsymbol{d}_0' | \boldsymbol{d}_1') = (\boldsymbol{b}'\boldsymbol{T} | \boldsymbol{b}'\boldsymbol{Y})$. Using $\boldsymbol{T}\boldsymbol{b} = \boldsymbol{0}$, $Q(\boldsymbol{U})$ can be written as

$$
\begin{aligned}
Q(\boldsymbol{U}) &= \begin{pmatrix} \boldsymbol{d}_0' & \boldsymbol{d}_1' \end{pmatrix} \begin{pmatrix} \boldsymbol{T}'\boldsymbol{D}_\pi\boldsymbol{T} & \boldsymbol{T}'\boldsymbol{D}_\pi\boldsymbol{Y} \\ \boldsymbol{Y}'\boldsymbol{D}_\pi\boldsymbol{T} & \boldsymbol{Y}'\boldsymbol{D}_\pi\boldsymbol{Y} \end{pmatrix}^- \begin{pmatrix} \boldsymbol{d}_0 \\ \boldsymbol{d}_1 \end{pmatrix} \\
&= \boldsymbol{d}_1'(\boldsymbol{Y}'\boldsymbol{D}_\pi\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{D}_\pi\boldsymbol{T}(\boldsymbol{T}'\boldsymbol{D}_\pi\boldsymbol{T})^-\boldsymbol{T}'\boldsymbol{D}_\pi\boldsymbol{Y})^-\boldsymbol{d}_1 \\
&= \boldsymbol{d}_1'\boldsymbol{W}_1^-\boldsymbol{d}_1
\end{aligned}
\tag{14}
$$

From (14) it can be seen how the matrix $\boldsymbol{T}$ influences $Q(\boldsymbol{U})$: although it has no contribution to the vector of deviates, it acts as a kind of correction on the matrix of the quadratic form. Generally speaking, the reason why this correction has to be carried out lies in the restrictions on the vector of deviates. Although $\boldsymbol{b}$ has $M$ elements, they can not all vary freely, because there are $F$ restrictions imposed by the likelihood equations and the elements of $\boldsymbol{b}$ sum to zero. This is also the background to why an $M$-vector with all elements equal to one must belong to the manifold of $\boldsymbol{T}$: this vector must be present to account for the restriction that the elements of $\boldsymbol{p}$ and $\pi$ sum to one. So the matrix of the quadratic form reflects the fact that the parameters are estimated from the data. In fact, Glas (1981) has shown that $\boldsymbol{W}$ is nothing else than the covariance matrix of $\boldsymbol{d}$, while $\boldsymbol{W}_1$ is the

covariance matrix of $d_1$ given $d_0$.

This section is concluded with a remark on the relation between $T$ and $Y$ and the appearance of generalized inverses in the (13) and (14). Using an over-parametrization for $T$ and imposing no restrictions on $Y$ is completely motivated by mathematical elegance. One could also adopt a full-rank parametrization of the model and add a unit vector to the associated matrix $Y$. Further, one could restrict $Y$ to have columns outside the manifold of $T$. In that case $U$ would be of full column rank and the generalized inverses in (13) and (14) could be replaced by proper inverses. However, removing columns from $U$ until it has full column rank neither changes the value of the statistic, nor its degrees of freedom. Therefore, the more elegant and less involved procedure for constructing the tests has been adopted here.

## Testing the Shape of the Item Characteristic Curves

In the framework of the Rasch model for dichotomous items, van den Wollenberg (1982) considered two tests: the $Q_1$-test, based on counts of the number of correct responses to the item in homogeneous score groups, and the $Q_2$-test, based on counts of simultaneous correct responses in homogeneous score groups. Van den Wollenberg (1982) presented a rationale which suggested that $Q_1$ has power against violation of the assumption of monotone increasing and parallel curves of item response functions, while $Q_2$ has power against multidimensionality. Various simulation studies (van den Wollenberg, 1979, 1982, Glas, 1981, 1988, 1989) corroborate this hypothesis. Glas (1988) has revised the $Q_1$- and $Q_2$- to the $R_{1c}$- and $R_{2c}$-test such that they fit the framework of

generalized Pearson tests and their asymptotic distribution could be derived. Glas and Verhelst (1995) have presented a further development of the $R_{1c}$, called the $S_i$-test, which has the same rationale as $R_{1c}$, but which focusses on specific items, hence the subscript $i$ .

In the present section, $R_{1c}$ and $S_i$ will be generalized to the GPCM, in the next section the same will be done for the $R_{2c}$-test. Further, a specialization of $R_{2c}$ will be presented which focusses on pairs of items. In both sections a theoretical framework will be presented for substantiating the claims with respect to the alternative models which are tested. Since it is a special case of the model considered here, this framework also applies to the Rasch model for dichotomous items. Therefore, this framework also provides a foundation for the claims with respect to the power of the $Q_1$-,$Q_2$-,$R_{1c}$- and $R_{2c}$-tests.

The tests considered here will be based on the difference between the counts of the numbers of persons belonging scoring $s$ and responding in category $h$ of item $i$, $M_{sih}$, with realization $m_{sih}$, and their CML expected values, $E(M_{gih}|\hat{\omega},\hat{\eta})$, that is, the expected value given the frequency distribution of the respondent's values of the minimal sufficient statistic for ability and the CML estimates of the item parameters. These differences will be denoted

$$d^*_{sih} = m_{sih} - E(M_{sih}|\hat{\omega},\hat{\eta}). \tag{15}$$

The expected frequencies are computed as

$$E(M_{sih}|\hat{\omega},\hat{\eta}) = \sum_{\{\boldsymbol{x}|x_{ih}=1,\boldsymbol{y}'\boldsymbol{r}=s\}} \hat{\pi}_{\boldsymbol{x}}, \tag{16}$$

where $\{\boldsymbol{x}|x_{ih}=1,\boldsymbol{y}'\boldsymbol{r}=s\}$ is the set of all possible response patterns with $x_{ih}=1$

and sum score $s$. Using (8), this expectation can be written as

$$E(M_{sih}|\hat{\omega},\hat{\eta}) = \frac{\omega_s}{\gamma_s}\sum_{\{\boldsymbol{x}|x_{ih}=1,\boldsymbol{y}'\boldsymbol{r}=s\}} \exp(-\boldsymbol{x}'\eta),$$

$$= \frac{\omega_s \varepsilon_{ih} \gamma^{(i)}_{s-r_{ih}}}{\gamma_s} \qquad (17)$$

where $\varepsilon_{ih} = \exp(\eta_{ih})$ and $\gamma^{(i)}_{s-r_{ih}}$ is a combinatorial function as defined in (7), only in this case response patterns are considered without the presence of item $i$, resulting in a sum score $s-r_{ih}$.

For any test of reasonable length, the number of deviates $d^*_{sih}$ is quite large, which results in two problems. Firstly, specific model violations are still hard to identify, and secondly, for certain combinations of $s$, $i$ and $h$ the expected frequencies $E(M_{sih}|\hat{\omega},\hat{\eta})$ may be too low to justify use of asymptotic theory.

---

Insert Table 2 about here

---

First, the $\boldsymbol{U}$-matrix of the test statistic will described using an example. Next, the rationale for building this specific matrix will be discussed. Continuing the example of Table 1, consider the matrix $\boldsymbol{U}$ of Table 2. The example is a test of three items with two, three and four response categories, respectively. The response patterns are the sufficient statistics, and, therefore, they are entered in $\boldsymbol{T}_1$. Further, the category score weights are given in the third row. The weighted sum score, which is the sufficient statistic for ability, ranges from one to eight. The sufficient statistics associated with the dummy score parameters $\omega_s$, $s = 1,...,8$, are entered into $\boldsymbol{T}_2$.

The statistic defined by $\boldsymbol{Y}$ will be based on a partition of the score range in three regions, consisting of the scores 1, 2, 3, and 4, the score 5, and the scores 6, 7 and 8, respectively and will be targeted at item 2. Therefore, $\boldsymbol{Y}$ consists of three groups of three column-vectors, each group is associated with one of the score regions, and in each group the rows consist of the possible response patterns on item 2, as far as the item response produces a response pattern with a sum score in the relevant the score range. Showing that $N\boldsymbol{U}'(\boldsymbol{p}-\hat{\pi})$ will produce the appropriate observed and expected frequencies proceeds as follows. Consider the last column of Table 1, where the elements of $\pi$ are listed and the first column of $\boldsymbol{Y}$ in Table 2. The inner product of these two vectors constitutes a sum over the probabilities of the response patterns with a sum score in the first region where the response to item 2 is in the zero category. Applying this principle to all columns of the $\boldsymbol{Y}$-matrix of Table 2, it can be verified that all differences $d_{g2h}$ can be produced by multiplying a column of the matrix $\boldsymbol{Y}$ with $N(\boldsymbol{p}-\hat{\pi})$. Notice, by the way, that all elements of the seventh column of $\boldsymbol{Y}$ are equal to zero. Therefore, this column will not produce any deviates and can be stricken without any consequence. In fact, also the three columns in $\boldsymbol{T}_1$ associated with item 2 can be removed, because they are contained in the linear manifold of $\boldsymbol{Y}$.

Next, the rationale for building this specific matrix and the rationale behind the test will be discussed. Above, $\boldsymbol{T}$ was introduced as a matrix of score functions related to the exponential family (9). Then this model was specialized to the GPCM, which serves as a null- hypothesis for the test. After estimation, it holds that $\boldsymbol{T}'(\boldsymbol{p}-\hat{\pi}) = \boldsymbol{0}$. As a model for the alternative hypothesis, consider a model where the item parameters differ with the score regions, that is, for every score region a different GPCM holds. This results in an exponential family model with a matrix $\boldsymbol{T}$ equal to the matrix $\boldsymbol{U}$ of the model test.

If the null-model holds, that is, if the same GPCM holds for all score regions, the elements of $\boldsymbol{U}(\boldsymbol{p}-\hat{\pi})$ will be close to zero, that is, the estimation equations for the alternative model are "almost" solved. However, if $\boldsymbol{U}(\boldsymbol{p}-\hat{\pi})$ departs significantly from zero, these equations are far from solved by the parameter estimates ensuing from the null-model, that is, adopting the alternative model may result in better model fit. Although from a statistical point of view this is perfectly feasible, in practice, the search for a better fitting model will not be in the direction of the alternative model, that is, in the direction of a model where every set of response patterns resulting a sum score in the same region will have its own GPCM. From a psychometric point of view the GPCM is far more parsimonious, and one may attempt to come to a better description of the observed and expected frequencies in the score regions by adjusting the score weights of the categories.

The final remark of this section concerns generalization of the item oriented $S_i$-test considered thus far to a global model test $R_{1C}$ that encompasses all items. This adaptation is almost trivial, for it consists of constructing a matrix $\boldsymbol{Y}$ that consists of all contrast vectors for all items as defined in Table 2. An interesting feature of this approach is that if all item contrasts are added to $\boldsymbol{Y}$, $\boldsymbol{T}_1$ can be removed from $\boldsymbol{U}$ altogether, because $\boldsymbol{T}_1$ is completely contained in the manifold of $\boldsymbol{Y}$. As a result, *woldW* has a block-diagonal form, and $Q(\boldsymbol{U})$ can be computed as a sum of $G$ squares.

## Testing Local Independence and Unidimensionality

Van den Wollenberg (1979, 1982) has shown that statistical tests for the Rasch model for dichotomous items based on comparing the observed and expected

counts of the number of correct scores on items in homogeneous score groups, such as the $Q_1$- and $R_{1C}$-test, are, in some instances, insensitive to violation of the axiom of unidimensionality. For instance, van den Wollenberg (1979) has proved that if a test is made up of two Rasch-homogeneous subtests that have equal item parameter vectors and the ability distributions associated with the two subtests are identical, test statistics based on the number of correct scores in score groups are insensitive to this model violation, regardless of the strength of the correlation between the two latent ability dimensions. Therefore, van den Wollenberg proposed a test statistic which is based on the following line of reasoning. Suppose unidimensionality is violated. If a subject's position on one dimension is fixed, the assumption of local stochastic independence requires that the association between the items vanishes. In the case of more than one dimension, however, the subject's position in the latent space is not sufficiently described by a unidimensional ability parameter and, as a consequence, the association between the responses to the items given the ability parameter will not vanish, that is, local independence is violated. Therefore, van den Wollenberg (1979, 1982) proposed a test, $Q_2$ that focusses on the observed and expected association between items. However, the asymptotic distribution of this test statistic has not been derived, though simulation studies (van den Wollenberg, 1979; Glas, 1981) support the conjecture that the test statistic has an approximate chi-square distribution. Practical application of the test has its limitations, because its computation requires CML parameter estimates at every possible score level. Glas (1988) has presented a revision of the test, called $R_{2C}$, that needs only one CML parameter estimation for its computation, and proved that the test statistic has an asymptotic chi-square distribution. In the present section, the above approach is applied to testing the GPCM. One of the main differences with the older version of $R_{2C}$

(Glas, 1988, 1989) will be that the test can also be computed per item pair.

A main feature that complicates testing local independence and unidimensionality is that the number of possible alternatives is quite large. Further, it must be stressed that local independence and unidimensionality are not the same thing. The models discussed by Kelderman and Rijkes (1994) and Glas (1992) are both multidimensional in the person parameters, but in their derivation local independence is definitely used. On the other hand, the model by Jannarone (1986), some models by Kelderman (1988) and the model by Verhelst and Glas (1993) lack the assumption of local independence but still are unidimensional. However, the thing all these models have in common is that analyzing data following these models using a unidimensional, locally independent Rasch model results in unexplained association between the items. Therefore, a general global statistic for testing the association between the items is presented, which is uninformative with respect to which model might work better. After presentation of this global test, some remarks will be made on how the testing against more specific alternatives might continue.

---

Insert Table 3 about here

---

Above, it was shown that the $\boldsymbol{U}$-matrix of a generalized Pearson statistic can be viewed as the $\boldsymbol{T}$-matrix of an alternative, more general model. The essence of the test presented here is enhancing the model of the null-hypothesis with parameters associated with pairs of items, to check to what extent these added parameters might contribute to explaining the association between items. Consider the

example of Table 3. This table has the same layout as Table 2, that is, it contains a matrix $\boldsymbol{U}$ consisting of a matrix $\boldsymbol{T}$ associated with the null-model and a matrix $\boldsymbol{Y}$ of the relevant contrasts or the score functions of the added parameters, whichever way one wants to look at it. The example of Table 3 concerns testing the association between item 1 and 2. The result of the product $\boldsymbol{N}\boldsymbol{Y}'\boldsymbol{p}$ must be the observed number of persons producing the simultaneous response pair $(x_{1h},x_{2k})$, for $h = 0,1$ and $k = 0,1,2$. In the same manner $\boldsymbol{N}\boldsymbol{Y}'\hat{\boldsymbol{\pi}}$ must produce its estimated expected value. Therefore, $\boldsymbol{Y}$ has six rows, which are associated with the pairs of categories $\{h,k\} = \{0,0\}, \{0,1\}, \{0,2\}, \{1,0\}, \{1,1\}, \{1,2\}$, respectively. For some pair $\{h,k\}$ the associated column of $\boldsymbol{Y}$ has as entries the product of $x_{1h}$ and $x_{2k}$, that is, the entry is one if the response pattern has a response on item 1 in category $h$ and a response in category $k$ on item 2. So essentially, the test is based on comparing the observed association in a two by six matrix produced by $\boldsymbol{N}\boldsymbol{Y}'\boldsymbol{p}$ with its expected value. If there turns out to be unexplained association, $\boldsymbol{N}\boldsymbol{Y}'(\boldsymbol{p}-\hat{\boldsymbol{\pi}})$ will significantly depart form zero and the model test will also be significant.

The final question that needs to be answered in this section is how to proceed if the global test for unidimensionality and local dependence is significant and the GPCM is rejected. Above, it was already mentioned that there are various alternatives to the GPCM. An important distinction between the alternative models is whether or not they can be estimated using CML. For models where CML applies, the testing procedure can be continued by entering the $\boldsymbol{T}$-matrix of the specific alternative of interest as a $\boldsymbol{U}$-matrix into the $Q(\boldsymbol{U})$-test. Specific alternatives may be unidimensional models by Jannarone (1986) and Kelderman (1984) lacking local independence and the multidimensional model by Kelderman and Rijkes (1994). A detailed description of this procedure is beyond the scope of

this chapter. A relevant model where CML is feasible is the multidimensional model by Glas (1992), where it is assumed that a test consists of a number of Rasch scales while ability has a multivariate normal distribution. This model can be estimated using MML. Glas and Verhelst (1995) have shown how the framework of generalized Pearson statistics can be adopted to IRT models extended with assumptions concerning the ability distribution. The essential requirement is that the extended model must have non-trivial, though not necessarily minimal sufficient statistics for the parameters, where non-trivial means that the aggregation level of the statistics must transcend the level of the mere response patterns themselves. Non-trivial sufficient statistics exist for the multidimensional model by Glas (1992), so the search for a fitting model can also be expanded into this direction. Also here a detailed description is beyond the scope of the present chapter.

## Some Simulated Examples

This chapter will be concluded with some simulation studies concerning the power of the tests. These studies do not have the pretention of being exhaustive, the purpose of this section is to give the reader some assistance in interpreting the outcomes of an analysis. The simulation studies will be focused on two topics, the power against improper specification of the score weights and the power against multi-dimensionality. All simulation studies were carried out with 100 replications. The sample size was 1000 respondents, augmenting the sample size did not produce any unexpected results, that is, the power of the tests grew larger. Therefore, the results for larger sample sizes will not be presented here.

---

Insert Table 4 about here

---

For the first example, concerning detection of improperly specified score weights, consider Table 4. The example consists of four items with four response categories each, that is $m_i = 3$, for $i = 1,...,4$. The item parameters and the score weights used for generating the data are shown in columns three to five. The weight of the zero category is fixed at zero for all studies to be presented. The means over 100 replications of the CML parameter estimates using the correct score weights are shown in column six, the mean estimated standard errors are in the next column. The estimation equations were solved using a Newton-Raphson algorithm. It can be seen that the algorithm performed properly: the parameter estimates all fall in a range of plus and minus two standard deviations around the true values of the parameters. For the next two studies, the score weights of the third item were changed from {1, 2, 3} to {2, 3, 4} and {3, 4, 5}, respectively. The resulting parameter estimates are displayed in the ninth and the last column of Table 4. It can be seen that all parameter estimates suffer from this improper specification, however the estimates of the parameters of the third item seem to suffer most. In Table 5 the results of testing model fit in these last two studies are summarized. The $S_i$- and $R_{1c}$-tests were computed using four score levels, that is, $G = 4$. For all reported studies a significance level of 5 % will be used. In the columns labeled "$S_i$", the mean value of the $S_i$ statistic over 100 replications is given, the columns labeled "Prob" and "%Sign." give the mean of the probability of the values of $S_i$ and the number of times that the test was significant in the 100 replications, respectively. The rows with the entry "$R_{1c}$" give the same information

for the computation of $R_{1C}$. Finally, the rows with the entry " $S_{ij}$ " give the mean of $S_{ij}$ and the mean of the probability values over all replications and all item pairs. Therefore, the column-entries under "%Sign." in these two rows refer to the percentage of significant outcomes in 600 model tests.

---

Insert Table 5 about here

---

It can be seen that in both studies item 3 is most often pinpointed as a misfit and, as expected, the number of significant outcomes of $S_i$ grows as the difference between the true and imputed score weights becomes larger. Imputing the score weights {4 5 6} for item 3 resulted in a significant result for all the S_i-tests for this item. However, also the number of significant results for the other items grows as the model violation for item 3 becomes more profound. The reasons for this phenomenon are that the estimates of the parameters of all items are affected by the model violation and that the total score as a criterium for forming homogeneous ability groups for computation of the test becomes more or less invalidated. Notice that the $S_{ij}$-test is not very sensitive to this model violation, though also here the number of significant outcomes grows with the importance of the model violation. The simulation studies reported in Table 6 follow the same lines as the previous ones, only here the number of response categories per item is varied. This time, one of the score weights of item 2 is changed dramatically, the true values {2, 5} are transformed to {1, 5} in a first study and {3, 5} in a second study. The parameters estimates using the true score weights are not reported in Table 6, they did not contain anything unexpected beyond the results already

shown in Table 4. The imputed score weights and the resulting mean estimates over 100 replications for the first study are shown in the columns six and seven, for the second study they are displayed in the columns ten and eleven. As expected, the model violations produce bias in the estimates. The columns labeled "$S_i$" give the percentage significant results for the $S_i$-test, at the bottom of the table the same is count is given for $R_{1c}$. It can be seen that item 2 is most often detected as a misfit. In the columns labeled "$S_{ij}$" the percentage significant outcomes is reported for all pairs of items $i,j$ where item $i$ is involved. So every entry in this column refers to 300 tests. It can be seen that $S_{ij}$ is far less sensitive to the model violation as $S_i$, but also here the percentage significant results is largest for item 2.

---

Insert Table 6 about here

---

Apart from supporting detection of misfitting items, $S_i$ also provides information on how to adjust score weights. It is beyond the scope of the present paper to develop a complete heuristic for this matter, however, an example will be given of how the information produced by the testing procedure can be applied to diagnostic purposes. In Table 7 information issuing from two replications of the simulation studies of Table 6 is presented, the first part of Table 7 relates to a study where item 2 has score weights {1, 5}, the second part of the table relates to a study where item 2 has score weights {3, 5}. So in the first analysis the index of category 1 is too low, in the second analysis the index of this category is too high. The $S_i$-test is based on the difference between observed and expected numbers

of responses on item categories in homogeneous score groups. In the two examples of Table 7 four score groups are formed, the first group has scores from 1 to 3, the second scores from 4 to 6, the third scores from 7 to 9 and the last group scores from 10 to 14. The score groups are formed in such a way that the numbers of respondents in each subgroup are approximately equal. In the table, the differences between observed and expected frequencies are divided by their standard deviations to produce so-called scaled deviates, which, approximately, have a standard normal distribution. The entry for $h = 2$ in the first score group is set equal to zero because it is not possible to simultaneously obtain a score less than or equal to 3 and respond in the second category of item 2. Comparing the rows of scaled deviates of item 2 in the two studies, it can be seen that, roughly speaking, the signs of the scaled deviates in both studies oppose. For instance, in the first study, there are less observations on the first category in the first score group than expected, while the opposite applies to the second study. Further, in the first study, there are more observations on the first category in the highest score group than expected, while in this group there are less responses on the second category than expected. Again, the opposite holds in the second study. So one may conclude that item 2 has too low a score weight in the first study and too high a weight in the second study. This, of course, complies with the manner in which the data were generated.

---

Insert Table 7 about here

---

The second set of simulations was aimed at the power of $S_{ij}$ against

multidimensionality. First consider the example of Table 8. In this example the items 1 and 4 related to one latent trait, while the items 2 and 3 related to another latent trait. Both ability variables had a standard normal distribution, the correlation between the variables was 0.50. Again 100 replications were made. The results are summarized in Table 8. Notice that the parameter estimates are systematically biased, in the sense that they shrink towards zero. For every item i, 100 $S_i$-tests and 300 $S_{ij}$-tests ($j = 1,..,4, i \neq j$) were computed, the percentages significant outcomes are summarized in the last two columns of Table 8. The percentage significant outcomes of $R_{1c}$ is given at the bottom of the table. Notice that all three statistics are sensitive to the model violation, only $S_i$ does a poor job for the first dichotomous item. With respect to S_{ij} it must be noticed that it did not matter whether the items $i$ and j related to the same dimension or not.

---

Insert Table 8 about here

---

The final set of simulations of this paper concerns the power of $S_i$ against multidimensionality. Above it was already mentioned that, for the case of dichotomous items, van den Wollenberg (1979) has proved that if a test is made up of two Rasch-homogeneous subtests that have identical item parameter vectors and ability distributions, test statistics based on the number of correct scores in score groups are insensitive to this model violation, regardless of the height of the correlation between the two latent ability dimensions. The present simulation study concerns the question whether the conditions identified by van den Wollenberg also apply to the case of polytomous items. The topic of this simulation study are

two subtests of four items each, the item parameters and score weights were the same as those reported for the four items of Table 8. The correlation between the two latent dimensions was fixed at 0.25 for the first 100 replications, 0.50 for the next 100 replications and 0.75 for the last 100 replications. The results are summarized in Table 9. The fifth, sixth and seventh column relate to the study with correlation equal to 0.25, the next three columns relate to the study with correlation 0.50 and the last three columns relate to the study with correlation 0.75. First of all it can be seen that the shrinkage in the parameter estimates lessens as the correlation becomes higher. Also the number of significant values of $S_{ij}$, $S_i$ and $R_{1c}$ reduces as the correlation becomes higher. However, for a correlation of 0.75, $S_{ij}$ is still significant half of the time, while the sensitivity of $S_i$ and $R_{1c}$ for the model violation has disappeared. Apparently, $S_{ij}$ is far better suited for detecting multidimensionality than $S_i$ and $R_{1c}$. Insensitivity to multidimensionality regardless of the height of the correlation between the two latent ability dimensions, however, does not hold for polytomous items.

---

Insert Table 9 about here

---

As a concluding remark in this section, it must be noticed that simulation studies unavoidably are to a large degree artificial. When analyzing real data, it will seldom be the case that one item violates one specific model assumption, while other items elicit only model conform responses. On the other hand, model violations may probably not be as profound as the ones studied here. In practical situations, it is advisable to start with an initial model that already accounts for possible

differences in discrimination between the items, rather than start with the basic partial credit model and try to adjust the hypotheses about the score weights by inspecting differences between observed and expected frequencies. Two possible initial models can be useful for this purpose: the nominal response model (Bock, 1972) and the OPLM (Verhelst and Glas, 1995). The nominal response model is equivalent with the GPCM defined in (2), except that in this case the scoring weights $r_{ih}$ are treated as unknown parameters to be estimated. It has already been mentioned above that the mathematical properties of the nominal response model are such that little progress has been made with respect to testing procedures for this model. However, Veldhuijzen (1995) and Verstralen (1995) have developed several heuristics based on this model to obtain initial values for the score functions in both the OPLM and the GPCM. The second approach is based on OPLM itself. In the OPLM a discrimination index is specified for every item and, therefore, the model can be viewed as a special case of the GPCM. For the OPLM the methods for adjusting hypothesis concerning discrimination indices using differences between observed and expected frequencies has been thoroughly worked out and works well in practice. Items that keep failing the OPLM can be further analyzed using the GPCM. Since the GPCM is quite flexible in the possibilities of modeling differences in discrimination between items and item categories, the main reason for failure of the GPCM might be multidimensionality. One of the obvious ways to proceed in case of lack of fit is to adopt the Rasch model with a multivariate distribution of ability (Glas, 1992) and replace the Rasch model with the GPCM. For this approach, two problems remain to be solved. Firstly, there must be available a practical heuristic for determining which items relate to the same latent distribution and, secondly, a testing procedure for the GPCM with a multivariate distribution of ability must be developed.

## References

Andersen, E.B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 34,* 42-50.

Andersen, E.B. (1973). *Conditional inference and models for measuring.* Unpublished dissertation, Mentalhygienisk Forskningsinstitut, Copenhagen.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Andersen, E.B. (1980). *Discrete statistical models with social science applications.* Amsterdam, North Holland.

Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B, 25,* 220-233.

Birnbaum, A. (1968). Some latent trait models. (hoofdstuk 17 in:) F.M. Lord & M.R. Novick, *Statistical theories of mental test scores.* Addison-Wesley: Reading (Mass.).

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm. *Psychometrika, 46,* 443-459.

Glas, C.A.W. (1981). *Het Raschmodel bij data in een onvolledig design [The Rasch model and missing data].* PSM-Progress reports, 81-1. Vakgroep PSM van de subfaculteit Psychologie: Utrecht.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525-546.

Glas, C.A.W. (1989). *Contributions to estimating and testing Rasch models.* Arnhem: Cito.

Glas, C.A.W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson, (Ed.), *Objective measurement: theory into practice, Vol. 1,* New Jersey: Ablex Publishing Corporation.

Glas, C.A.W., and Verhelst, N.D. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635-659.

Glas, C.A.W., and Verhelst, N.D. (1995). Tests of fit for polytomous Rasch models. In G.H.Fischer & I.W.Molenaar (eds.). *Rasch models. Their foundations, recent developments and applications.* New York: Springer.

Haberman, S.J. (1974). *The analysis of frequency data.* Chicago: University of Chicago Press.

Jannarone, R.J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51,* 357-373.

Kelderman, H. (1984). Loglinear RM tests. *Psychometrika, 49,* 223-245.

Kelderman, H. and Rijkes, C.P.M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59,* 149-176.

Martin Löf, P. (1973). *Statistika Modeller. Anteckningar från seminarier Lasåret 1969-1970, utarbetade av Rolf Sunberg. Obetydligt ändrat nytryck, oktober 1973.* Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Neyman, J., and Scott, E.L. (1948). Consistent estimates, based on partially consistent observations. *Econometrica, 16,* 1-32.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1961). *On the general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* 321-333. Berkeley: University of California Press.

Van den Wollenberg, A.L. (1979). *The Rasch model and time limit tests.* Nijmegen: Studentenpers.

Van den Wollenberg, A.L. (1982). Two new tests for the Rasch model. *Psychometrika, 47,* 123-140.

Veldhuijzen, N.H. (1995). *A Heuristic Procedure for Suggesting an Appropriate Scoring Function for Polytomous Items with Ordered Response Categories.* Measurement and Research Department Reports, 95-1. Cito: Arnhem.

Verhelst, N.D., and Glas, C.A.W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58,* 395-415.

Verhelst, N.D., and Glas, C.A.W. (1995). The generalized one parameter model: OPLM. In: G.H.Fischer & I.W.Molenaar (eds.). *Rasch models: their foundations, recent developments and applications.* New York: Springer.

Verhelst, N.D., Glas, C.A.W. and Verstralen, H.H.F.M. (1995). *OPLM: computer program and manual.* Arnhem: Cito.

Verstralen, H.H.F.M. (1995). *Optimal Integer Category Weights in the OPLM and GPCM.* Measurement and Research Department Reports, 95-2. Cito: Arnhem.

Wilson, M. and Masters, G.N.(1993). The partial credit model and null categories. *Psychometrika, 58,* 85-99.

Table 1
An example of the matrix $T$

| item | 1 | 2 | 3 | | |
|---|---|---|---|---|---|
| cat. | 01 | 012 | 0123 | | |
| weight | 12 | 013 | 0123 | | |
| score | | | | score 12345678 | probability |
| 1 | 10 | 100 | 1000 | 10000000 | $\pi(10, 100, 1000)$ |
| 2 | 10 | 100 | 0100 | 01000000 | $\pi(10, 100, 0100)$ |
| 2 | 10 | 010 | 1000 | 01000000 | $\pi(10, 010, 1000)$ |
| 2 | 01 | 100 | 1000 | 01000000 | $\pi(01, 100, 1000)$ |
| 3 | 01 | 010 | 1000 | 00100000 | $\pi(01, 010, 1000)$ |
| 3 | 10 | 100 | 0010 | 00100000 | $\pi(10, 100, 0010)$ |
| 3 | 10 | 010 | 0100 | 00100000 | $\pi(10, 010, 0100)$ |
| 3 | 01 | 100 | 0100 | 00100000 | $\pi(01, 100, 0100)$ |
| 4 | 10 | 100 | 0001 | 00010000 | $\pi(10, 100, 0001)$ |
| 4 | 10 | 001 | 1000 | 00010000 | $\pi(10, 001, 1000)$ |
| 4 | 01 | 010 | 0100 | 00010000 | $\pi(01, 010, 0100)$ |
| 4 | 10 | 010 | 0010 | 00010000 | $\pi(10, 010, 0010)$ |
| 4 | 01 | 100 | 0010 | 00010000 | $\pi(01, 100, 0010)$ |
| 5 | 10 | 010 | 0001 | 00001000 | $\pi(10, 010, 0001)$ |
| 5 | 10 | 001 | 0100 | 00001000 | $\pi(10, 001, 0100)$ |
| 5 | 01 | 100 | 0001 | 00001000 | $\pi(01, 100, 0001)$ |
| 5 | 01 | 001 | 1000 | 00001000 | $\pi(01, 001, 1000)$ |
| 5 | 01 | 010 | 0010 | 00001000 | $\pi(01, 010, 0010)$ |
| 6 | 01 | 010 | 0001 | 00000100 | $\pi(01, 010, 0001)$ |
| 6 | 01 | 001 | 0100 | 00000100 | $\pi(01, 001, 0100)$ |
| 6 | 10 | 001 | 0010 | 00000100 | $\pi(10, 001, 0010)$ |
| 7 | 10 | 001 | 0001 | 00000010 | $\pi(10, 001, 0001)$ |
| 7 | 01 | 001 | 0010 | 00000010 | $\pi(01, 001, 0010)$ |
| 8 | 01 | 001 | 0001 | 00000001 | $\pi(01, 001, 0001)$ |

Table 2
An Example of the Matrix $U$
for Testing the ICC's of Item 2
(the entries left blank are equal to zero)

| | $T_1$ | | | $T_2$ | | $Y$ | |
|---|---|---|---|---|---|---|---|
| item | 1 | 2 | 3 | | | | |
| cat. | 01 | 012 | 0123 | | | | |
| weight | 12 | 013 | 0123 | | | | |
| score | | | | score 12345678 | | | |
| 1 | 10 | 100 | 1000 | 10000000 | 100 | | |
| 2 | 10 | 100 | 0100 | 01000000 | 100 | | |
| 2 | 10 | 010 | 1000 | 01000000 | 010 | | |
| 2 | 01 | 100 | 1000 | 01000000 | 100 | | |
| 3 | 01 | 010 | 1000 | 00100000 | 010 | | |
| 3 | 10 | 100 | 0010 | 00100000 | 100 | | |
| 3 | 10 | 010 | 0100 | 00100000 | 010 | | |
| 3 | 01 | 100 | 0100 | 00100000 | 100 | | |
| 4 | 10 | 100 | 0001 | 00010000 | 100 | | |
| 4 | 10 | 001 | 1000 | 00010000 | 001 | | |
| 4 | 01 | 010 | 0100 | 00010000 | 010 | | |
| 4 | 10 | 010 | 0010 | 00010000 | 010 | | |
| 5 | 01 | 100 | 0001 | 00001000 | | 100 | |
| 5 | 01 | 001 | 1000 | 00001000 | | 001 | |
| 5 | 01 | 010 | 0010 | 00001000 | | 010 | |
| 6 | 01 | 010 | 0001 | 00000100 | | | 010 |
| 6 | 01 | 001 | 0100 | 00000100 | | | 001 |
| 6 | 10 | 001 | 0010 | 00000100 | | | 001 |
| 7 | 10 | 001 | 0001 | 00000010 | | | 001 |
| 7 | 01 | 001 | 0010 | 00000010 | | | 001 |
| 8 | 01 | 001 | 0001 | 00000001 | | | 001 |

Table 3
An Example of the Matrix $U$
for item 1 and 2

| | | $T_1$ | | $T_2$ | $Y$ |
|---|---|---|---|---|---|
| item | 1 | 2 | 3 | | |
| cat. | 01 | 012 | 0123 | | |
| weight | 12 | 013 | 0123 | | |
| | | | | score | |
| score | | | | 12345678 | |
| 1 | 10 | 100 | 1000 | 10000000 | 100000 |
| 2 | 10 | 100 | 0100 | 01000000 | 100000 |
| 2 | 10 | 010 | 1000 | 01000000 | 010000 |
| 2 | 01 | 100 | 1000 | 01000000 | 000100 |
| 3 | 01 | 010 | 1000 | 00100000 | 000010 |
| 3 | 10 | 100 | 0010 | 00100000 | 100000 |
| 3 | 10 | 010 | 0100 | 00100000 | 010000 |
| 3 | 01 | 100 | 0100 | 00100000 | 000100 |
| 4 | 10 | 100 | 0001 | 00010000 | 100000 |
| 4 | 10 | 001 | 1000 | 00010000 | 001000 |
| 4 | 01 | 010 | 0100 | 00010000 | 000010 |
| 4 | 10 | 010 | 0010 | 00010000 | 010000 |
| 4 | 01 | 100 | 0010 | 00010000 | 000100 |
| 5 | 10 | 010 | 0001 | 00001000 | 010000 |
| 5 | 10 | 001 | 0100 | 00001000 | 001000 |
| 5 | 01 | 100 | 0001 | 00001000 | 000100 |
| 5 | 01 | 001 | 1000 | 00001000 | 000001 |
| 5 | 01 | 010 | 0010 | 00001000 | 000010 |
| 6 | 01 | 010 | 0001 | 00000100 | 000010 |
| 6 | 01 | 001 | 0100 | 00000100 | 000001 |
| 6 | 10 | 001 | 0010 | 00000100 | 001000 |
| 7 | 10 | 001 | 0001 | 00000010 | 001000 |
| 7 | 01 | 001 | 0010 | 00000010 | 000001 |
| 8 | 01 | 001 | 0001 | 00000001 | 000001 |

Table 4
Parameter Estimates Using Correct and Incorrect Discrimination Indices.

| | | True Values | | Study 1 | | | Study 2 | | Study 3 | |
|---|---|---|---|---|---|---|---|---|---|---|
| i | h | $\beta$ | $\eta$ | r | $\hat{\eta}$ | $se(\hat{\eta})$ | r | $\hat{\eta}$ | r | $\hat{\eta}$ |
| 1 | 1 | -1.50 | -1.50 | 1 | -1.424 | .114 | 1 | -1.293 | 1 | -1.231 |
| | 2 | .00 | -1.50 | 2 | -1.345 | .134 | 2 | -1.282 | 2 | -1.225 |
| | 3 | 1.50 | .00 | 3 | .034 | .170 | 3 | .272 | 3 | .060 |
| 2 | 1 | -.67 | -.67 | 2 | -.594 | .134 | 2 | -.558 | 2 | -.371 |
| | 2 | .00 | -.67 | 5 | -.430 | .237 | 5 | -.486 | 5 | -.449 |
| | 3 | .67 | .00 | 7 | .462 | .290 | 7 | -.086 | 7 | .089 |
| 3 | 1 | -.67 | -.67 | 1 | -.710 | .105 | 2 | -1.006 | 3 | -1.387 |
| | 2 | .00 | -.67 | 2 | -.758 | .126 | 3 | -.717 | 4 | -1.268 |
| | 3 | .67 | .00 | 3 | .122 | .158 | 4 | -.099 | 5 | -.722 |
| 4 | 1 | -1.00 | -1.00 | 1 | -1.220 | .099 | 1 | -.930 | 1 | -.664 |
| | 2 | .00 | -1.00 | 3 | -1.026 | .089 | 3 | -.903 | 3 | -.857 |
| | 3 | 1.00 | .00 | 4 | .000 | — | 4 | .000 | 4 | .000 |

Table 5
Testing Model Fit

| Item | Weights | | | | $S_i$ | DF | Prob | %Sign. |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 2 | 3 | 10.805 | 9 | .411 | 12 |
| 2 | 0 | 2 | 5 | 7 | 8.643 | 9 | .545 | 8 |
| 3 | 0 | 2 | 3 | 4 | 16.122 | 9 | .197 | 39 |
| 4 | 0 | 1 | 3 | 4 | 11.054 | 9 | .371 | 10 |
| $R_{1c}$ | | | | | 46.626 | 30 | .138 | 57 |
| $S_{ij}$ | | | | | 9.624 | 9 | .501 | 47 |
| Item | Weights | | | | $S_i$ | DF | Prob | %Sign. |
| 1 | 0 | 1 | 2 | 3 | 14.406 | 9 | .301 | 33 |
| 2 | 0 | 2 | 5 | 7 | 12.412 | 9 | .319 | 25 |
| 3 | 0 | 3 | 4 | 5 | 30.407 | 9 | .033 | 87 |
| 4 | 0 | 1 | 3 | 4 | 16.235 | 9 | .204 | 38 |
| $R_{1c}$ | | | | | 73.461 | 30 | .026 | 93 |
| $S_{ij}$ | | | | | 12.387 | 9 | .371 | 33 |

Table 6
Parameter Estimates and Model Tests

| | | | True Values | | Study 1 | | | | Study 2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| i | h | r | $\beta$ | $\eta$ | r | $\hat{\eta}$ | $S_i$ | $S_{ij}$ | r | $\hat{\eta}$ | $S_i$ | $S_{ij}$ |
| 1 | 1 | 1 | .00 | .00 | 1 | .093 | 6 | 12 | 1 | -.020 | 4 | 7 |
| 2 | 1 | 2 | -.67 | -.67 | 1 | -.192 | | | 3 | -.900 | | |
| | 2 | 5 | .67 | .00 | 5 | .488 | 74 | 19 | 5 | -.224 | 48 | 7 |
| 3 | 1 | 1 | -.67 | -.67 | 2 | -.692 | | | 3 | -.491 | | |
| | 2 | 2 | .67 | .00 | 3 | -.111 | 4 | 5 | 4 | .147 | 4 | 0 |
| 4 | 1 | 1 | -1.00 | -1.00 | 1 | -1.020 | | | 1 | -.970 | | |
| | 2 | 3 | .00 | -1.00 | 3 | -.997 | | | 3 | -.911 | | |
| | 3 | 4 | 1.00 | .00 | 4 | .000 | 2 | 3 | 4 | .000 | 2 | 1 |
| $R_{1c}$ | | | | | | | 66 | | | | 52 | |

Table 7
Patterns of Scaled Deviates

| i | h | r | Group-1 1 to 3 | Group-2 4 to 6 | Group-3 7 to 9 | Group-4 10 to 14 | SS | $S_i$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | .328 | -.016 | -.578 | .463 | 3.385 | 3.149 |
| 2 | 1 | 1 | -2.569 | .448 | 1.579 | 1.483 | 13.783 | |
| | 2 | 5 | .000 | .920 | .621 | -1.114 | 4.643 | 17.472 |
| 3 | 1 | 1 | .759 | -.296 | .140 | -.401 | 3.433 | |
| | 2 | 2 | .085 | -.191 | -.416 | .386 | 2.971 | 5.986 |
| 4 | 1 | 1 | .399 | -.198 | .135 | -.564 | 2.948 | |
| | 2 | 3 | .344 | -.139 | -.078 | -.006 | 2.633 | |
| | 3 | 4 | .100 | -.185 | -.422 | .319 | 1.986 | 7.126 |
| 1 | 1 | 1 | -.462 | -.050 | .710 | -.511 | 3.706 | 3.165 |
| 2 | 1 | 3 | 1.624 | .843 | -.086 | -1.560 | 8.497 | |
| | 2 | 5 | .000 | -1.282 | -1.195 | 1.446 | 6.761 | 13.091 |
| 3 | 1 | 1 | -.561 | .735 | -.069 | -.062 | 3.010 | |
| | 2 | 2 | -.279 | -.255 | .558 | -.117 | 3.240 | 5.414 |
| 4 | 1 | 1 | -.561 | .014 | -.231 | .700 | 3.222 | |
| | 2 | 3 | -.461 | .321 | .605 | -.320 | 3.579 | |
| | 3 | 4 | .100 | -.269 | .292 | -.043 | 1.700 | 7.186 |

Table 8
Parameter Estimates and Model Tests
with Multidimensional Data

| i | h | r | Dim | $\beta$ | $\eta$ | $\hat{\eta}$ | $S_i$ | $S_{ij}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 3 | 1 | .00 | .00 | -.004 | 5 | 84 |
| 2 | 1 | 2 | | -.50 | -.50 | -.193 | | |
| | 2 | 4 | 2 | .50 | .00 | .014 | 65 | 76 |
| 3 | 1 | 2 | | -.50 | -.50 | -.238 | | |
| | 2 | 3 | 2 | .50 | .00 | .099 | 14 | 52 |
| 4 | 1 | 1 | | -1.00 | -1.00 | -.642 | | |
| | 2 | 3 | | .00 | -1.00 | -.582 | | |
| | 3 | 4 | 1 | 1.00 | .00 | .000 | 64 | 51 |
| $R_{1c}$ | | | | | | | 97 | |

Table 9
Parameter Estimates and Model tests
for Two Equal Subtests

| | | | | Study 1 | | | Study 2 | | | Study 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| i | h | r | Dim | $\hat{\eta}$ | $S_i$ | $S_{ij}$ | $\hat{\eta}$ | $S_i$ | $S_{ij}$ | $\hat{\eta}$ | $S_i$ | $S_{ij}$ |
| 1 | 1 | 3 | 1 | -.260 | 13 | 100 | -.090 | 8 | 99 | -.094 | 5 | 63 |
| 2 | 1 | 2 | | .007 | | | -.149 | | | -.311 | | |
| | 2 | 4 | 1 | -.065 | 34 | 100 | -.152 | 13 | 99 | -.087 | 5 | 57 |
| 3 | 1 | 2 | | -.231 | | | -.363 | | | -.423 | | |
| | 2 | 3 | 1 | -.087 | 3 | 99 | .054 | 15 | 99 | -.107 | 4 | 41 |
| 4 | 1 | 1 | | -.611 | | | -.625 | | | -.945 | | |
| | 2 | 3 | | -.680 | | | -.733 | | | -1.144 | | |
| | 3 | 4 | 1 | -.211 | 17 | 99 | -.192 | 10 | 96 | -.268 | 3 | 41 |
| 5 | 1 | 3 | 2 | .079 | 13 | 100 | -.128 | 5 | 99 | -.174 | 6 | 58 |
| 6 | 1 | 2 | | -.004 | | | -.312 | | | -.396 | | |
| | 2 | 4 | 2 | .000 | 35 | 100 | -.303 | 20 | 97 | -.126 | 9 | 56 |
| 7 | 1 | 2 | | -.088 | | | -.367 | | | -.344 | | |
| | 2 | 3 | 2 | .088 | 8 | 99 | -.006 | 7 | 89 | -.116 | 5 | 42 |
| 8 | 1 | 1 | | -.494 | | | -.494 | | | -.639 | | |
| | 2 | 3 | | -.430 | | | -.701 | | | -.857 | | |
| | 3 | 4 | 2 | .000 | 13 | 99 | .000 | 9 | 93 | .000 | 5 | 40 |
| $R_{1c}$ | | | | | 56 | | | 32 | | | 6 | |

**Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.**

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.

**U.S. Department of Education**
*Office of Educational Research and Improvement (OERI)*
*National Library of Education (NLE)*
*Educational Resources Information Center (ERIC)*

# NOTICE

## REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").